2018

A Topological Approach to Understanding Location-Based Data

McAbee, Carson C.; Wakefield, Max D.; Roth, John D.; Scrofani, James W.

IEEE

McAbee, Carson, et al. "A Topological Approach to Understanding Location-Based Data." 2018 52nd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2018.
http://hdl.handle.net/10945/63116

A Topological Approach to Understanding Location-Based Data

Carson McAbee, Max D. Wakefield, John D. Roth, and James W. Scrofani

***Abstract*—Location-based services have seen a boon in data production recently which has simultaneously stoked the research community to better understand this type of information. Traditional methods in analyzing such data require significant *a priori* understanding of the organization of the data. We submit that the nascent field of topological data analysis (TDA) may be able to contribute new insights to analysis of such data without the aforementioned requirement. To this end, we propose two novel methods of embedding such data in order to leverage the expressive power of TDA. To demonstrate its effectiveness we apply the embeddings to maritime automated information system data.**

***Index Terms*—Machine Learning, Topological Data Analysis**

I. Introduction

With the advent of location-based services (LBS) an increasing amount of location data is being generated. Consequently, this unique dataset catalyzes interest in mining that data for relevant patterns for a variety of moving objects [1], [2]. Out of this new research effort applications such as transportation system efficiency [3], vehicle congestion reduction [4], taxi dispatching, and location-based advertising [5], among others, will benefit.

Increasingly complex data [5], [6] require increasingly complex solutions. On one hand, traditional statistical methods are powerful, but only if the nature of the data is understood *a priori*. On the other hand, neural networks are powerful when data are more opaque, but require a sufficient quantity and quality of training data. Therefore, drawing meaning from big data requires either a valid assumption on known relationships or sufficient training data and computational resources.

Recently, an alternative to understanding data has emerged: persistent homology [7]–[9]. At its most elemental realization it is a mathematical query about the topology of the data (i.e., shape and organization). Meaning can then be extracted from such knowledge and follow-on analyses can be better informed [6]. Persistent homology requires no presuppositions on the nature of the data and has mature computational algorithms rendering it practical in the face of modern data [6], [8], [10], [11]. Indeed, the engineering community has begun to realize its potential by bringing it to bear on problems such as medical imaging [12], pulmonary diagnoses [13], and cellular network optimization [14].

We propose the application of persistent homology to the study of location-based data. To this end we present two novel embedding methods. We then demonstrate their efficacy on synthetic and real-world maritime automated information system (AIS) location data.

II. A Brief Primer on Persistent Homology

The central focus of persistent homology is to detect various geometric and shape properties of data where distance between points is defined. Persistent homology uses abstract simplicial complexes which are a set of subsets of a set, usually finite, that is closed under subset. From some data or some model one constructs a sequence of simplicial complexes $\Delta(*)$: $\Delta(\epsilon_1) \subseteq \Delta(\epsilon_2) \subseteq \cdots \subseteq \Delta(\epsilon_k) \subseteq \cdots$ (cf. Figure 1), which are unions and gluings of higher dimensional complexes. The next ingredient is a algebraic analytic in the form of a vector space or module called homology $H_i(\Delta(\mathbb{K}))$ that records essential shape information such as how many connected components ($i = 0$), holes ($i = 1$), or enclosed spaces ($i = 2$) are in the complex $\Delta(\mathbb{K})$. The homology modules are the quotient of the kernel by the image in the following complex:

$$\cdots \xrightarrow{\partial_{i+1}} C_i(\Delta(\epsilon_j)) \xrightarrow{\partial_i} C_{i-1}(\Delta(\epsilon_j)) \xrightarrow{\partial_{i-1}} \cdots$$

where the chain groups $C_i(\Delta(\epsilon_j))$ are the free abelian groups generated by the dimension i simplicies (subsets in the simplicial complex of size $i+1$) and the maps ∂_i take a simplex $\sigma = \{x_1, \ldots, x_{i+1}\}$ to the alternating sum of its deletions

$$\partial_i(\sigma) = \sum_{a=1}^{i+1} (-1)^a \sigma \backslash \{x_a\}.$$

Hence the i^{th} homology module of $\Delta(\epsilon_j)$ is

$$H_i(\Delta(\epsilon_j)) = \frac{\ker(\partial_i)}{\operatorname{Im}(\partial_{i+1})}.$$

Finally using the natural inclusion maps from the sequence of simplicial complexes the *persistent homology* of the sequence is the image of these maps

$$H_i^{j\to k}\big(\Delta(\star)\big) = \operatorname{Im}\left[H_i(\Delta(\epsilon_j)) \to H_i(\Delta(\epsilon_k))\right].$$

The non-zero homological elements in $H_i^{j\to k}\big(\Delta(\star)\big)$ are said to persist and indicate essential geometric phenomena of the sequence $\Delta(\star)$. These non-zero elements can be visualized as horizontal lines, as in Figure 1, starting at coordinate j where the element first appears and stopping at k where the element becomes zero. The union of all these horizontal lines is called a barcode [15]. Topological features that persist for some significant range are considered salient whereas those which are more transitory may be considered a kind of topological noise.

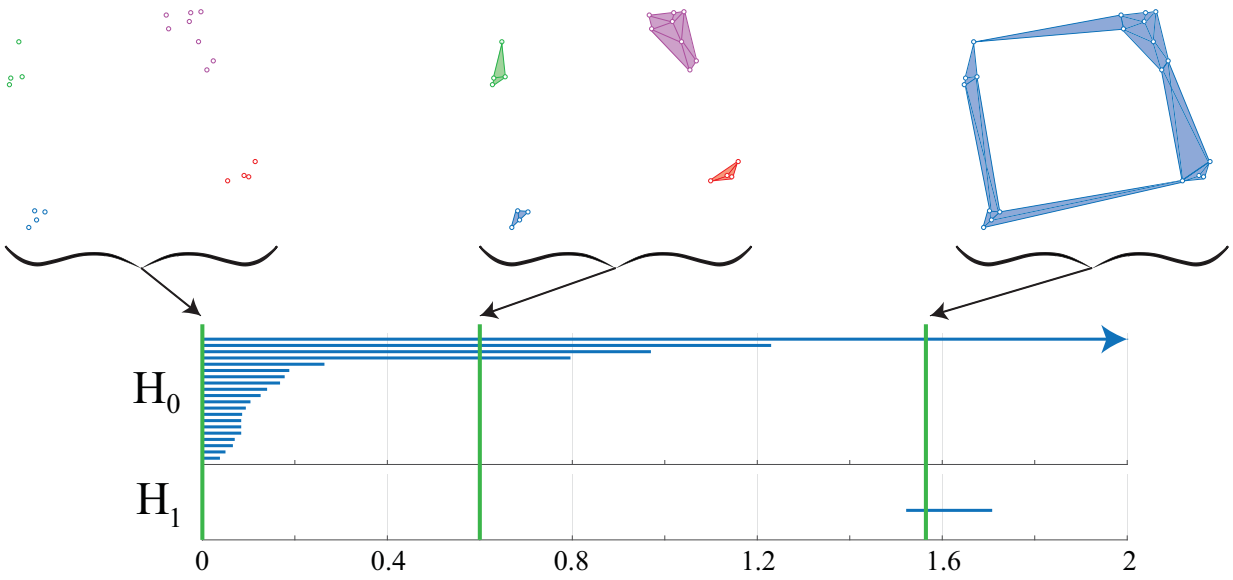

Fig. 1. An example filtration over 20 data points is presented with its corresponding H_0 & H_1 features. The barcode at three stages of filtration is presented with each corresponding simplicial complex. The H_0 features are seen to decay per their respective density as discussed in Section III-A. The single H_1 feature is representative of the hole seen in the right-most complex. This hole disappears once the filtration value becomes large enough to fill in that hole.

III. Approach

Among the many candidates of spatio-temporal data which may benefit from the applications of persistent homology, we choose AIS data which encodes location-based data $\hat{X} \in \mathbb{R}^2$ of maritime vessels due to its wide availability and usefulness in ensuring safety at sea. AIS data also include a variety of other information such as speed and heading [16].

Initially, we simulate AIS data representing two vessel types. Commercial vessels hold more true to course and change speed and direction less often. On the other hand, fishing vessels are more erratic and change direction and speed frequently. Each vessel is guided by a Markov model which embodies the characteristics of the aforementioned behavior along a voyage in which the vessel visits two ports enroute to the final destination.

A. Persistence of the Difference to an Ideal Track

To leverage H_0 features (i.e., connected components), an embedding function f_0 is applied to the location data $\hat{X}$. Here, the distance function compares each datum in the AIS data with a null model X that represents the optimal vessel track via

$$f_0(i) = \hat{X}_k - X_k. \tag{1}$$

Implicitly, f_0 computes the difference in speed and heading with the null model X. The ideal track X is computed by taking the start and end points along the voyage, connecting them with a straight line, and dividing it into equal time segments. The resulting H_0 persistence, is computed on $f_0(k)$, $\forall k$ and mapped to a representative barcode. An exponential decay $e^{-\alpha\epsilon}$ is then fit to the ends of the barcode, ultimately projecting the data onto a one-dimensional space in the form of a decay coefficient α. This function is observed to decay much more quickly when the ideal and actual track difference is small due to the tightly grouped set of points projected in the neighborhood of the origin. This tight cluster of points causes them to connect quickly. Conversely, if the ideal and actual track difference is large, the points are more scattered and connect slower during the filtration ϵ thus creating slower decay.

The resulting labeled data are presented in Figure 2 where it is clear that each of the vessel types are generally separated in α. In one sense, this method can also be seen as a type of topological distance estimator, since the denser the points are in the embedded space the faster they will decay. We then prescribe a Rayleigh distribution to the resulting distributions of α in order to derive a appropriate threshold for classification

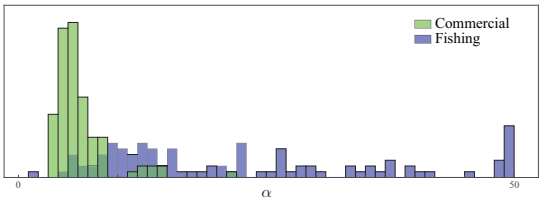

Fig. 2. The H_0 feature persistence is computed on the track $\hat{X}$ of each vessel. An exponential decay profile is then fit to the ends of each bar in the barcode effectively mapping the AIS track data to a decay coefficient presented on the ordinate.

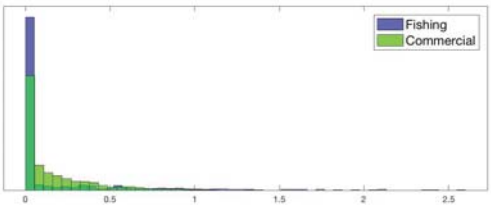

Fig. 3. The H_1 feature persistence is computed on the track $\hat{X}$ of each vessel. The sum of these persistence ranges is then computed mapping the AIS track data to the parameter presented on the ordinate.

via the Neyman-Pearson lemma.

B. Persistence of the Difference to a Local Average

Here we seek to leverage a different embedding for vessel classification using both H_0 and H_1 homology but without needing to define a true track. The goal is to perform vessel classification based on the hole behavior in the data. To this end we propose a different embedding function f_1 that computes the difference between a vessel's speed and heading $\{s_k, h_k\}$ and the mean speed and heading $\{\mu_{s,k}, \mu_{h,k}\}$ over some local number of points

$$f_1 = \{s_k, h_k\} - \{\mu_{s,k}, \mu_{h,k}\}. \tag{2}$$

The differences are then cast into $\mathbb{R}^2$ where the H_0 and the H_1 persistent homology is then calculated. The sum of all of the ranges of the H_i persistence is then computed as a single parameter β_i. Thus, similar to the previous embedding function f_0, f_1 assimilates the data into a one-dimensional space.

For the commercial vessels the f_1 data points are clustered around the origin tightly since their course and speed rarely change. However for fishing vessels the data points are much more spread out uniformly around the origin. Hence there will be much less H_0 persistence for the commercial vessel when compared to the fishing vessels.

The idea behind studying β_1 is to detect more subtle geometric patterns in location-based data. For example, vessels that are not piloted by either professional commercial captains or professional fisherman will exhibit phenomena that have both commercial and fishing characteristics, we call these mixed class vessels. The goal for β_1 is to distinguish the mixed class from both the commercial and the fishing. Again, under the invariant f_1 the commercial data points are tightly clustered around the origin and the fishing data points are evenly spread out away from the origin. The mixed class data points are pushed away from the origin, but tend to not be spread out as far from the origin as the fishing vessels due to a lack of drastic changes in course and speed. Hence, this mixed class lends itself well to exhibiting dimension 1 cycles (aka loops) and therefore H_1.

The invariant β_1 does not lend itself well to distinguishing fishing from commercial. The result is shown in Figure 3 where separation in β_1 is apparent but not drastic. Here we find that vessels that do not have high variation in heading and speed (i.e., commercial vessels) exhibit more H_1 features. The opposite is true of the fishing vessels. This phenomenon can be explained with a similar approach to the one used previously. A wide dispersion of points will lend itself more to the possibility of these H_1 holes. Alternatively, tightly clustered points will present less opportunity for such topological features.

IV. Results

In order to quantify the efficacy of f_0 and f_1 for object classification we examine them with both synthetic and real-world AIS data.

A. Synthetic Data

As discussed previously, synthetic data are generated with two Markov models: one for speed and one for heading. The speed Markov model has three states which correspond to the maintenance of speed, increasing speed, and slowing speed. Limits are placed on the speed to prevent unrealistically high or negative speed. The heading Markov model has two states that correspond to maintaining track or deviating from track. The amount of deviation from track or from current speed is chosen randomly. The state transition probabilities were tuned to highly favor the states corresponding to maintenance of speed and heading in the case of the commercial tracks. This bias was slightly weakened for fishing vessel tracks while still maintaining an overall propensity towards the maintenance states.

To test the embedding functions a cross-validation approach is adopted such that 10% of the data are used for training and the remaining 90% are used for testing. This approach is repeated some number of trials until a convergence of results is achieved. Overall, 1000 tracks per vessel type are used.

During the initial training phase an optimal threshold is established given the training data via the Neyman-Pearson lemma. The threshold is derived either assuming Rayleigh distributed data (in the case of f_0) or uniformly distributed data (in the case of f_1). Once the threshold is obtained binary classification of vessels is performed with the specific goal of identifying commercial vessels. For identifying commercial vessels using f_0 the resulting precision and recall are 0.7 and 0.9 respectively.

With this synthetic data we created a third vessel class which is an approximate average of the Markov chain transition matrices of the commercial and fishing vessels. It is difficult to distinguish this mixed class from the commercial and fishing vessels using the H_0 persistent homology model. However, a good distinguishing invariant using f_1 with H_1 is available. To support this claim we note that the average over a 1000 trials of the sums of H_1 bar code lengths for each vessel type is 0.1535 for commercial, 0.1482 for fishing, and 0.5687 for this third vessel type. The much larger average persistence of H_1 features in the mixed class data suggest that one dimensional cycles are distinguishing topological features.

Looking at all three classes using f_1 with β_0 (H_0) we get precision 0.90 and recall 0.79 for detecting commercial

vessels, precision 0.99 and recall 0.91 for detecting fishing vessels, and precision 0.53 and recall 0.53 for the mixed class. This validates our intuition that H_0 persistence will do well distinguishing the commercial and fishing vessels but not the mixed class.

In fact, distinguishing the mixed class is more difficult, but significant results are still achieved with β_1 (cf. Figure 4). Using this β_1 we get precision 0.77 and recall 0.56 for detecting commercial, precision 0.82 and recall 0.59 for detecting fishing vessels, and precision 0.54 and recall 0.79 for the mixed class. These promising results lead us to continue research on H_1 persistence for more delicate analysis.

B. Real-World Data

Having demonstrated efficacy on controlled synthetic data, f_0 and f_1 are then applied to real-world AIS data. Ground truth is obtained by a navigational status field used in AIS which indicates whether or not a vessel is underway using engine or engaged in fishing [16]. From a corpus of data the longest ten tracks are selected for each vessel type. We refer hereafter to underway using engine as a commercial vessel and engaged in fishing as a fishing vessel. As no ground truth can be obtained for a mixed vessel type, only classification of commercial and fishing vessels is considered here.

The cross-validation approach is used as with the synthetic data with the one exception of having less tracks available to train and test with. Additionally, since only one datum was used for training each distribution, only a simple mean between the commercial and fishing samples was used to set a threshold.

Again, binary classification with the goal of identifying commercial vessels was performed. Using f_0, the resulting precision and recall are 0.9 and 0.6 respectively. Using f_1 with β_0 the resulting precision and recall are 0.61 and 0.91 respectively for the commercial vessels and 0.95 and 0.74 for the fishing class. As predicted, using f_1 with β_1 we get poor results at 0.63 precision and 0.56 recall for distinguishing commercial from fishing.

C. Discussion

The discriminatory power of topological data analysis is highlighted by our invariants f_0 and f_1. For f_0, the embedding function aside, the H_0 computation can be seen as a kind of agglomerative clustering approach where the rate of clustering is what matters. In another way, this can be seen as a novel kind of point field density estimator. Here, object behavior is embodied in the *density* of the projected points. Significant detection rates are realized using this method both with real and synthetic data. Additionally, this embedding appears to be robust to sparse levels of data.

Good results are obtained on synthetic data using f_1 indicating the usefulness of higher-dimensional topological information. Interestingly, f_1 requires less information than f_0 since f_0 uses an optimal track. Unfortunately, one cannot necessarily obtain in real time under real-world situations an optimal track. Also, the results are decent and f_1 still has some power to distinguish more delicate vessel types not fitting within the clear setting of commercial or fishing vessels.

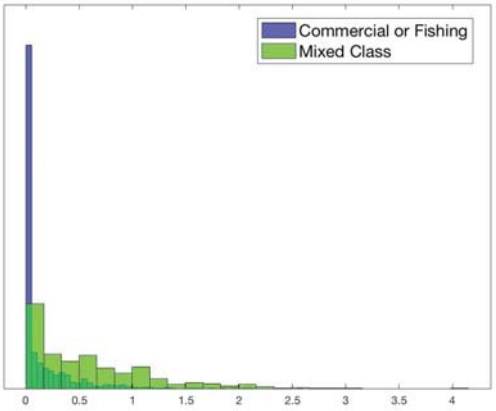

Fig. 4. The H_1 persistence for the mixed class vessels is presented here alongside commercial and fishing vessels. The longer tail of the mixed class distribution indicates higher average H_1 persistence with f_1.

V. Summary

In summary, we have presented a novel application of persistent homology for analysis of location-based data. The first embedding f_0 leverages the zeroth-dimensional topological features H_0 for classification by way of density estimation. The second embedding leverages the first-dimensional topological features H_1. Both embedding functions were similar in that they reduced the feature space to a single dimension. It was observed that tightly packed points (i.e., commercial vessels) clustered quickly and if the density of points was sufficiently different (i.e., commercial and fishing vessels) good classification results were possible using H_0 topology. However, if point field densities were close, as seen with the mixed class vessel type, H_0 was not as effective. It was shown how higher dimensional features in H_1 were then able to still perform this difficult classification problem.

References

[1] B. Barz, E. Rodner, Y. Guanche-Garcia, and J. Denzler, "Detecting regions of maximal divergence for spatio-temporal anomaly detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2018.

[2] H. Nguyen, W. Liu, and F. Chen, "Discovering congestion propagation patterns in spatio-temporal traffic data," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 169–180, Apr. 2016.

[3] S. Liu, Y. Yue, and R. Krishnan, "Non-myopic adaptive route planning in uncertain congestion environments," *IEEE Trans. Knowledge Data Engineering*, vol. 27, no. 9, pp. 2438–2451, Sep. 2015.

[4] J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, "Planning bike lanes based on sharing-bikes trajectories," in *Proc. 23rd ACM SIGKDD Int. Conf. Kownledge Discovery Data Mining*, 2017, pp. 1377–1386.

[5] Y. Ding, Y. Li, X. Zhou, Z. Huang, S. You, and J. Luo, "Sampling big trajectory data for traversal trajectory aggregate query," *IEEE Trans. Big Data*, Aug. 2018.

[6] A. Phinyomark, E. Ibáñez Marcelo, and G. Petri, "Resting-state fMRI functional connectivity: Big data preprocessing pipelines and topological data analysis," *IEEE Trans. Big Data*, vol. 3, no. 4, pp. 415–428, Sep. 2017.

[7] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification," *Discrete Comput. Geom.*, vol. 28, no. 4, pp. 511–533, 2002, discrete and computational geometry and graph drawing (Columbia, SC, 2001). [Online]. Available: http://dx.doi.org/10.1007/s00454-002-2885-2

[8] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete Comput. Geom.*, vol. 33, no. 2, pp. 249–274, 2005. [Online]. Available: http://dx.doi.org/10.1007/s00454-004-1146-y

[9] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc. (N.S.)*, vol. 46, no. 2, pp. 255–308, 2009. [Online]. Available: http://dx.doi.org/10.1090/S0273-0979-09-01249-X

[10] V. De Silva and G. E. Carlsson, "Topological estimation using witness complexes." *SPBG*, vol. 4, pp. 157–166, 2004.

[11] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "JavaPlex: A research software package for persistent (co)homology," in *Proc. ICMS 2014*, ser. Lecture Notes in Computer Science 8592, H. Hong and C. Yap, Eds., 2014, pp. 129–136, software available at http://appliedtopology.github.io/javaplex/.

[12] M. Lee and D. Han, "Voronoi tessellation based interpolation method for Wi-Fi radio map construction," *IEEE Commun. Letters*, vol. 16, no. 3, pp. 404–407, 2012.

[13] S. Emrani, T. Gentimis, and H. Krim, "Persistent homology of delay embeddings and its application to wheeze detection," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 459–463, Apr. 2014.

[14] A. Vergne, L. Decreusefond, and P. Martins, "Simplicial homology for future cellular networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 8, pp. 1712–1725, Aug. 2015.

[15] R. Ghrist, "Barcodes: The persistent topology of data," *Bull. Amer. Math. Soc.*, vol. 45, pp. 61–75, 2008.

[16] U.S. Coast Guard Navigation Center (n.d.), "Automatic information system overview." [Online]. Available: https://navcen.uscg.gov/?pageName=AISmain